Applied Statistical Regression
AS 2012 – Week 06

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, October 29, 2012

Multiple Linear Regression

We use linear modeling for a multiple predictor regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + E$$

- there are now p predictors
- the problem cannot be visualized in a scatterplot
- there will be n observations of response and predictors
- goal: estimating the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ from the data

IMPORTANT: simple linear regression of the response on each of the predictors does not equal multiple regression, where *all predictors are used simultanously*.

Comparing Hierachical Models

Idea: Correctly comparing two multiple linear regression models when the smaller has >1 predictor less than the bigger.

Where and why do we need this?

- for the 3 pollution variables in the mortality data.
- soon also for the so-called factor/dummy variables.

Idea: We compare the residual sum of squares (RSS):

Big model: $y = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q + \beta_{q+1} x_{q+1} + ... + \beta_p x_p$

Small model: $y = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q$

The big model must contain all the predictors from the small model, else they are not hierarchical and the test does not apply.

Applied Statistical Regression
AS 2012 – Week 06

The Global F-Test

***Idea**: is there any relation between response and predictors?*

This is another hierachical model comparison. The full model is tested against a small model with only the intercept, but without any predictors.

We are testing the null $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ against the alternative $H_A : \beta_j \neq 0$ for at least one predictor x_j. This test is again based on comparing the RSS:

$$F = \frac{n-(p+1)}{p} \cdot \frac{RSS_{Small} - RSS_{Big}}{RSS_{Big}} \sim F_{p,n-(p+1)}$$

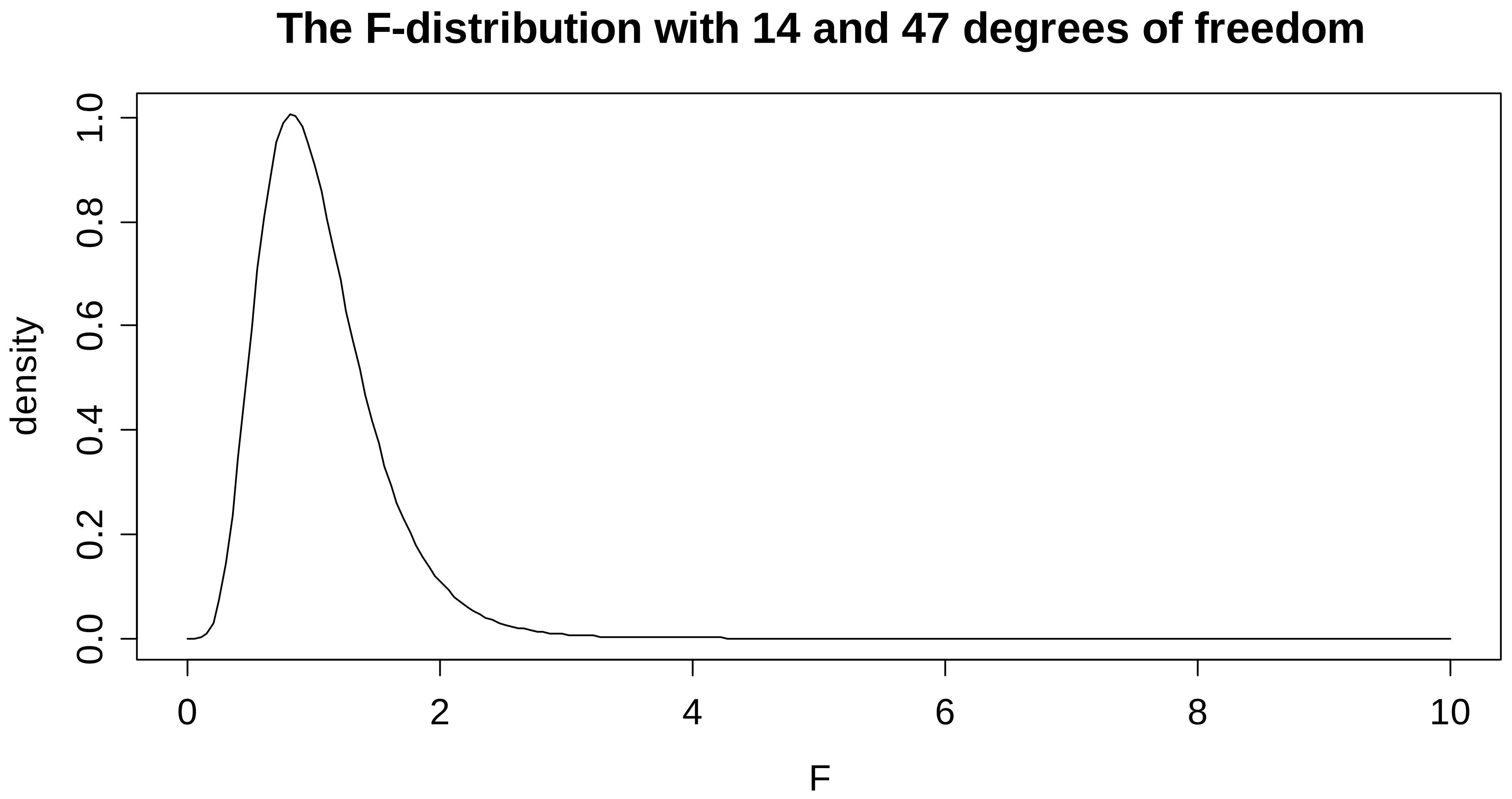
→ **Test statistic and p-value are shown in the R summary!**

Reading R-Output

```
> summary(fit.orig)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1496.4915   572.7205   2.613  0.01224 *
JanTemp       -2.4479     0.8808  -2.779  0.00798 **
...
Dens          11.9490    16.1836   0.738  0.46423
NonWhite     326.6757    62.9092   5.193 5.09e-06 ***
WhiteCollar -146.3477   112.5510  -1.300  0.20028
...
---
Residual standard error: 34.23 on 44 degrees of freedom
Multiple R-squared: 0.7719,Adjusted R-squared: 0.6994
F-statistic: 10.64 on 14 and 44 DF,  p-value: 6.508e-10
```

Note: due to space constraints, this is only a part of the output!

Density Function of the F-distribution



Prediction

The regression equation can be employed to predict the response value for any given predictor configuration.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{.1} + \hat{\beta}_2 x_{.2} + \ldots + \hat{\beta}_p x_{.p}$$

Note:
This can be a predictor configuration that was not part of the original data. For example a (new) city, for which only the predictors are known, but the mortality is not.

Be careful:
Only interpolation, i.e. prediction within the range of observed y-values works well, extrapolation yields non-reliable results.

Prediction in R

We can use the regression fit for predicting new observations. The syntax is as follows

```
> fit.big <- lm(Mortality ~ ., data=mt)
> dat       <- data.frame(JanTemp=..., ...)
> predict(fit.big, newdata=dat)
1 932.488
```

The x-values need to be provided in a data frame. The variable (column) names need to be identical to the predictor names. Of course, all predictors need to be present.

Then, it is simply applying the `predict()`-procedure.

Confidence- and Prediction Interval

The confidence interval for the fitted value and the prediction interval for future observation also exist in multiple regression.

a) 95%-CI for the fitted value $E[y \mid x]$

```
> predict(fit, newdata=dat, "confidence")
```

b) 95%-PI for a future observation $\hat{y}$:

```
> predict(fit, newdata=dat, "prediction")
```

- The visualization of these intervals is no longer possible in the case of multiple regression
- It is possible to write explicit formulae for the intervals using the matrix notation. We omit them here.

Reading R-Output

```
> summary(fit.orig)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1496.4915   572.7205   2.613  0.01224 *
JanTemp       -2.4479     0.8808  -2.779  0.00798 **
...
Dens          11.9490    16.1836   0.738  0.46423
NonWhite     326.6757    62.9092   5.193 5.09e-06 ***
WhiteCollar -146.3477   112.5510  -1.300  0.20028
...
---
Residual standard error: 34.23 on 44 degrees of freedom
Multiple R-squared: 0.7719, Adjusted R-squared: 0.6994
F-statistic: 10.64 on 14 and 44 DF,  p-value: 6.508e-10
```

Note: due to space constraints, this is only a part of the output!

Versatility of Multiple Linear Regression

Despite that we are using linear models only, we have a versatile and powerful tool. While the response is always a continuous variable, different predictor types are allowed:

- **Continuous Predictors**
  Default case, e.g. *temperature*, *distance*, *pH-value*, …
- **Transformed Predictors**
  For example: $log(x),\ sqrt(x),\ arcsin(\sqrt{x}),...$
- **Powers**
  We can also use: $x^{-1},\ x^2,\ x^3,\ ...$
- **Categorical Predictors**
  Often used: *sex*, *day of week*, *political party*, …

Applied Statistical Regression
AS 2012 – Week 06

First-Aid Transformations

This is a guideline as to how the variables in a regression can and should be transformed. The recommendation is to always apply these except if there are strong reasons against. From a practical viewpoint, they stabilize variance and improve the fit.

Absolute values, concentrations, right-skewed variables:
log-transformation: $x' = \log(x)$ and also $y' = \log(y)$

Count variables:
square-root transformation: $x' = \sqrt{x}$, maybe also $x' = \log(x)$

Proportions:
arcsine transformation: $x' = \sin^{-1}\left(\sqrt{x}\right)$

First-Aid Transformations

Example: Zurich Airport Data

Both the *predictor ATM* and the *response Pax* are count variables that only take positive values. They are due to a FAT. Because of the easier interpretation, we prefer to take logarithms here.

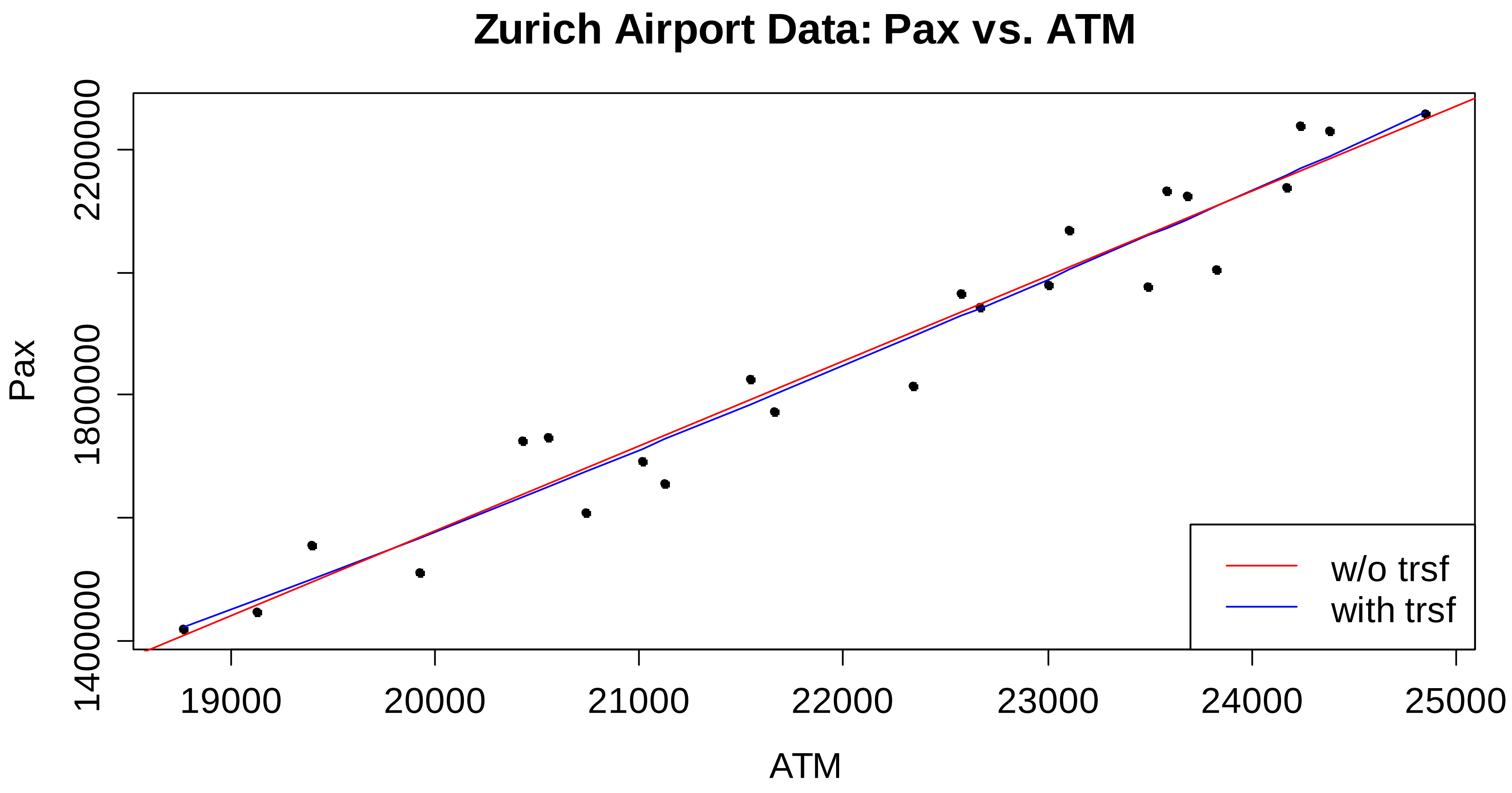
$$ATM' = \log(ATM) \qquad Pax' = \log(Pax)$$

The R code is as follows:

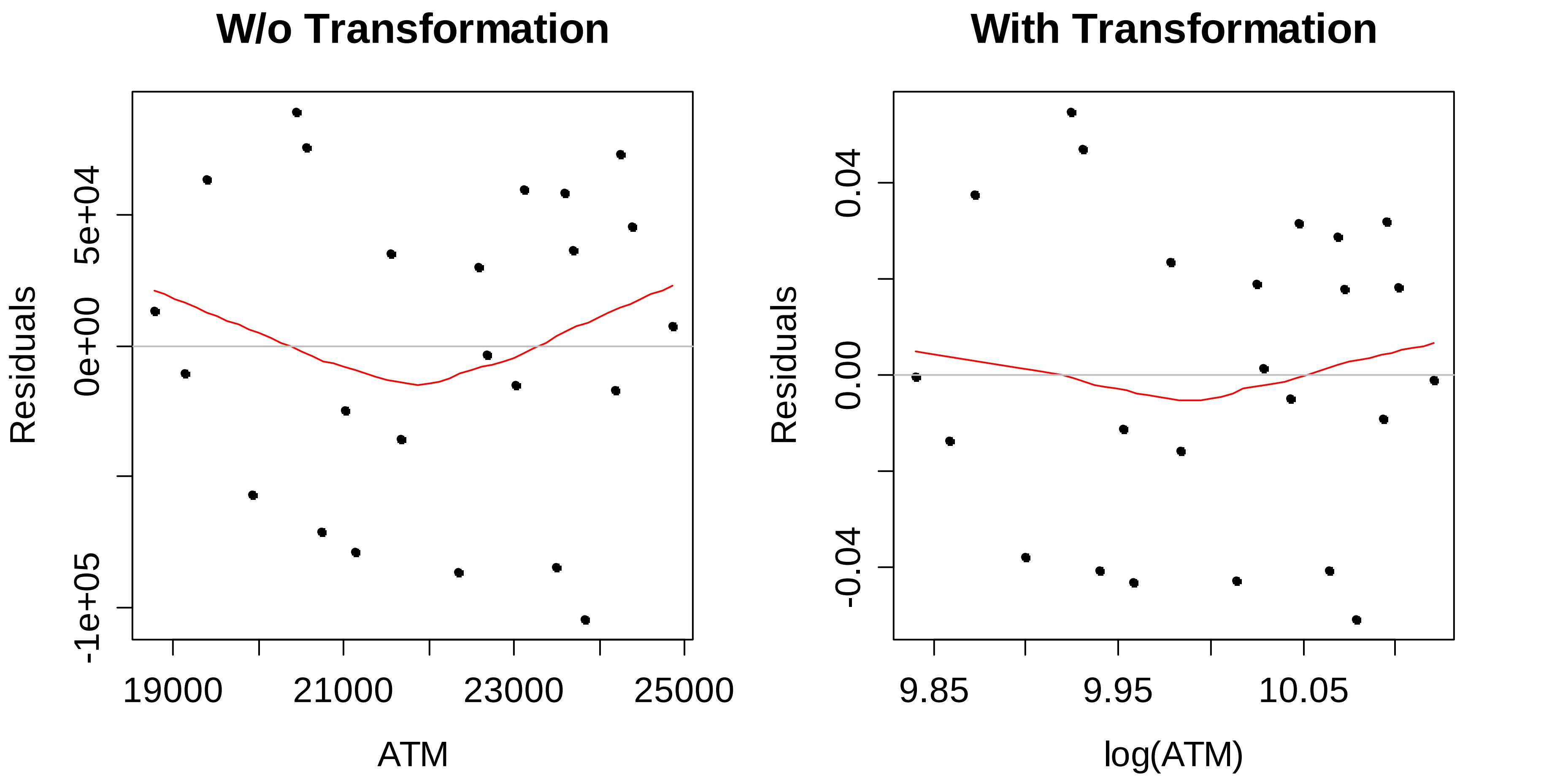
```
> fit.log <- lm(log(Pax) ~ log(ATM), data=...)
```

The fit is no longer a straight line but a curve. And there is no longer a linear increase in Pax with rising ATM, but...

Straight Line vs. log-log Fit



Comparison of Residuals vs. Predictor



Conclusions for Zurich Airport Data

The assumptions on the error are better fulfilled and we obtain smaller residuals after the log-log transformation. Thus, this is the more accurate model.

```
> lm(log(Pax) ~ log(ATM), data=...)
(Intercept)      log(ATM)
     -2.116          1.655
```

The relation is: $y = \exp(-2.116) \cdot x^{1.655}$, resp. $Pax = 0.120 \cdot ATM^{1.655}$

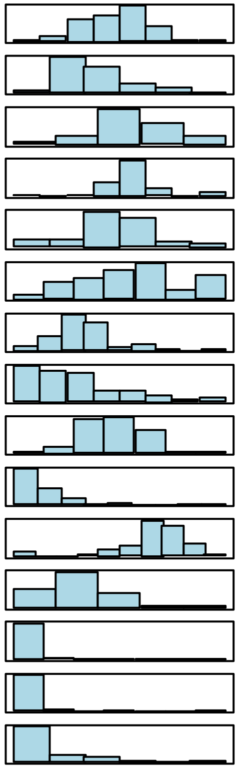
Thus, if ATM increases by 1%, then Pax increases by 1.655%. This is due to bigger airplanes used and higher seat load factor during busy months.

Applied Statistical Regression
AS 2012 – Week 06

FAT for the Mortality Data

The following variable transformations are recommended:

```
> str(mortality)
'data.frame':    59 obs. of  16 variables:
 $ Mortality  : num  922 998 962 ...
 $ JanTemp    : num  27 23 29 45 ...
 $ JulyTemp   : num  71 72 74 79 ...
 $ RelHum     : num  59 57 54 56 ...
 $ Rain       : num  36 35 44 47 ...
 $ Educ       : num  11.4 11 9.8 ...
 $ Dens       : num  3243 4281   ...
 $ NonWhite   : num  8.8 3.5 0.8 ...
 $ WhiteCollar: num  42.6 50.7   ...
 $ Pop        : num  660328 83588...
 $ House      : num  3.34 3.14   ...
 $ Income     : num  29560 31458 ...
 $ HC         : num  21 8 6 18   ...
 $ NOx        : num  15 10 6 8   ...
 $ SO2        : num  59 39 33 24 ...
```



The Effect of Variable Transformations

Under non-linear variable transformations (i.e. *log*, *sqrt* or *arcsin*), most results change: *coefficients*, *fitted values*, *tests & p-values*.

```
> anova(fit.trsf.big, fit.trsf.small)

Analysis of Variance Table

Model 1: Mortality ~ JanTemp + JulyTemp + RelHum + Rain +
         Educ + Dens + NonWhite + WhiteCollar + Pop +
         House + Income + log(HC) + log(NOx) + log(SO2)
Model 2: Mortality ~ JanTemp + JulyTemp + RelHum + Rain +
         Educ + Dens + NonWhite + WhiteCollar + Pop +
         House + Income

  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     45 53917
2     48 65672 -3    -11755 3.2703 0.02967 *
```

Linear Variable Transformations

Example: American Automobile Dataset

```
> head(mtcars, 10)
                   mpg cyl  disp  hp drat    wt  qsec vs am
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0
Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0
Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0
```

→ Fuel consumption is measured in *mpg* instead of *l/100km* and displacement in *cubic inches* but not *ccm*. **Can we convert?**

Linear Variable Transformations

Changing units, i.e. all linear variable transformations are allowed. While the regression coefficients change, fitted values, test stats, p-values and model diagnostics remain the very same!

Since the results are easier to read, it has proven very important to use well-readable and natural units for regression analysis

```
mile    <- 1.609344
gallon  <- 3.78541178
l.100km <- 100/(dat$mpg*mile/gallon)

inch    <- 2.54
ccm     <- dat$disp*(2.54^3)
```

Categorical Predictors

The canonical case in linear regression are *continuous predictor variables* such as for example:

→ *temperature, distance, pressure, velocity, ...*

While in linear regression, we cannot have categorical response, it is perfectly valid to have *categorical predictors*:

→ *yes/no, sex (m/f), type (a/b/c), shift (day/evening/night), ...*

Such categorical predictors are often also called **factor variables**. In a linear regression, each level of such a variable is encoded by a dummy variable, so that $(\ell - 1)$ degrees of freedom are spent.

Example: Binary Categorical Variable

The lathe (*in German: Drehbank*) **dataset:**

- y lifetime of a cutting tool in a turning machine
- x_1 speed of the machine in rpm
- x_2 tool type A or B

Dummy variable encoding:

$$x_2 = \begin{cases} 0 & tool \;\; type \;\; A \\ 1 & tool \;\; type \;\; B \end{cases}$$

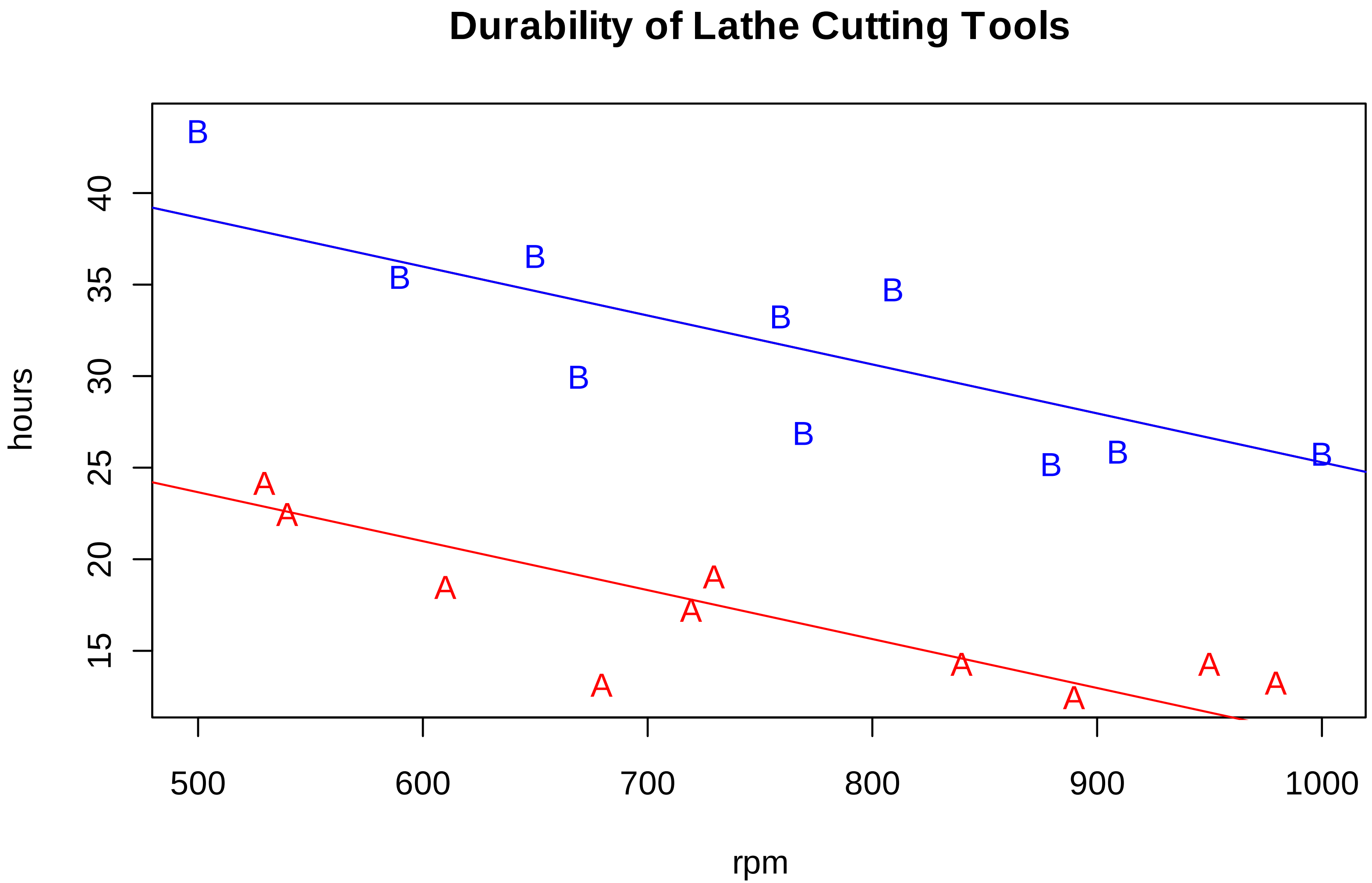
Interpretation of the Model

→ **see blackboard…**

```
> summary(lm(hours ~ rpm + tool, data = lathe))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.98560    3.51038  10.536 7.16e-09 ***
rpm         -0.02661    0.00452  -5.887 1.79e-05 ***
toolB       15.00425    1.35967  11.035 3.59e-09 ***
---
Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-squared: 0.9003,  Adjusted R-squared: 0.8886
F-statistic: 76.75 on 2 and 17 DF,   p-value: 3.086e-09
```

Applied Statistical Regression
AS 2012 – Week 06

The Dummy Variable Fit



A Model with Interactions

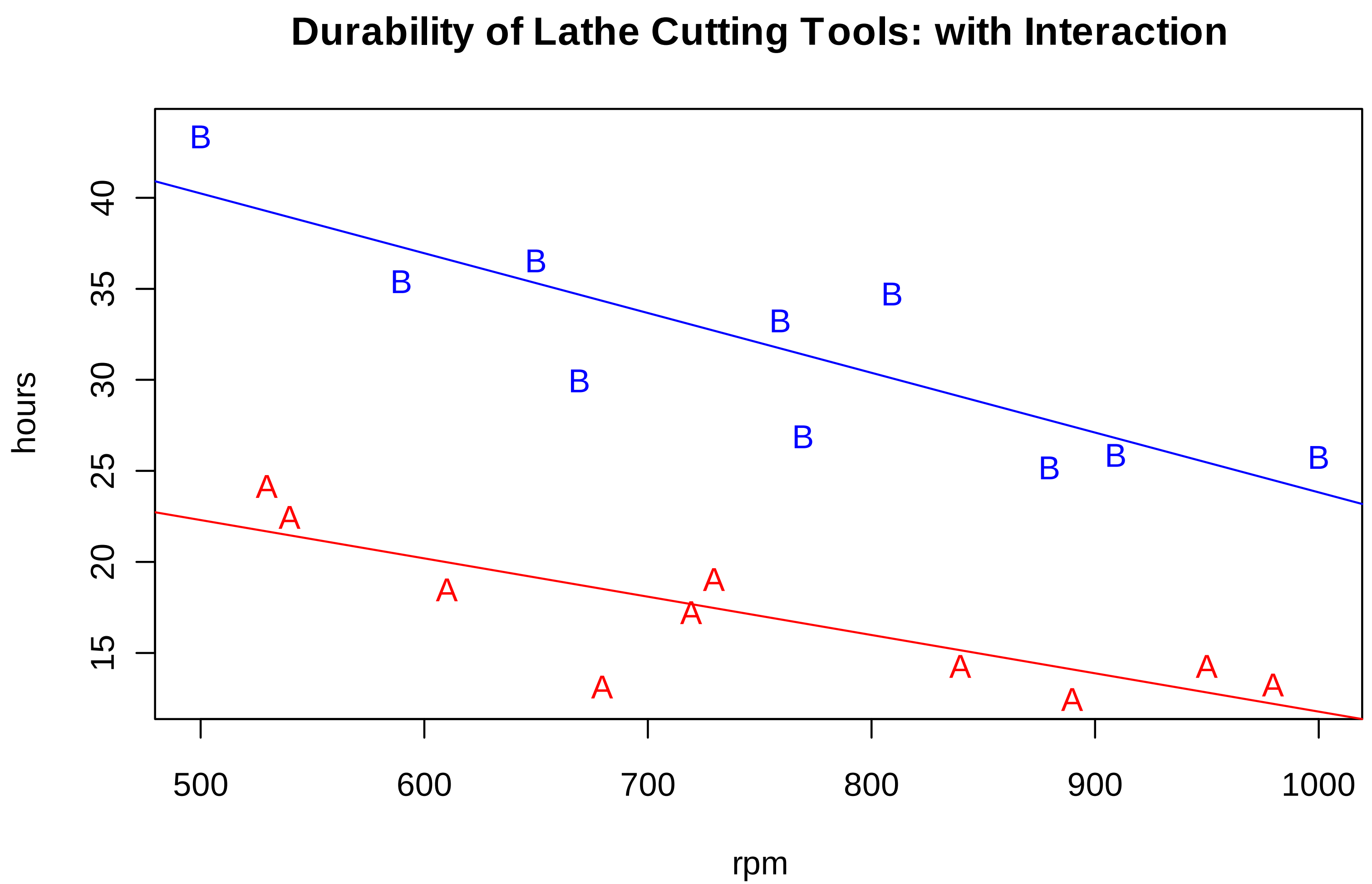
Question: do the slopes need to be identical?

→ with the appropriate model, the answer is no!

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + E$$

→ **see blackboard for model interpretation…**

Different Slopes for the Regression Lines



Summary Output

```
> summary(lm(hours ~ rpm * tool, data = lathe))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.774760   4.633472   7.073 2.63e-06 ***
rpm         -0.020970   0.006074  -3.452  0.00328 **
toolB       23.970593   6.768973   3.541  0.00272 **
rpm:toolB   -0.011944   0.008842  -1.351  0.19553
---
Residual standard error: 2.968 on 16 degrees of freedom
Multiple R-squared: 0.9105,  Adjusted R-squared: 0.8937
F-statistic: 54.25 on 3 and 16 DF,  p-value: 1.319e-08
```

How Complex the Model Needs to Be?

Question 1: do we need different slopes for the two lines?

$$H_0 : \beta_3 = 0 \text{ against } H_A : \beta_3 \neq 0$$

→ no, see individual test for the interaction term on previous slide!

Question 2: is there any difference altogether?

$$H_0 : \beta_2 = \beta_3 = 0 \text{ against } H_A : \beta_2 \neq 0 \ \ and\,/\,or \ \ \beta_3 \neq 0$$

→ this is a hierarchical model comparison
→ we try to exclude interaction and dummy variable together

R offers convenient functionality for this test, see next slide!

Testing the Tool Type Variable

Hierarchical model comparison with anova():

```
> fit.small <- lm(hours ~ rpm, data=lathe)
> fit.big <- lm(hours ~ rpm * tool, data=lathe)
> anova(fit.small, fit.big)
Model 1: hours ~ rpm
Model 2: hours ~ rpm * tool
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     18 1282.08
2     16  140.98  2    1141.1 64.755 2.137e-08 ***
```

→ The bigger model, i.e. making a distinction between the tools, is significantly better. The main effect is enough, though.

Categorical Input with More Than 2 Levels

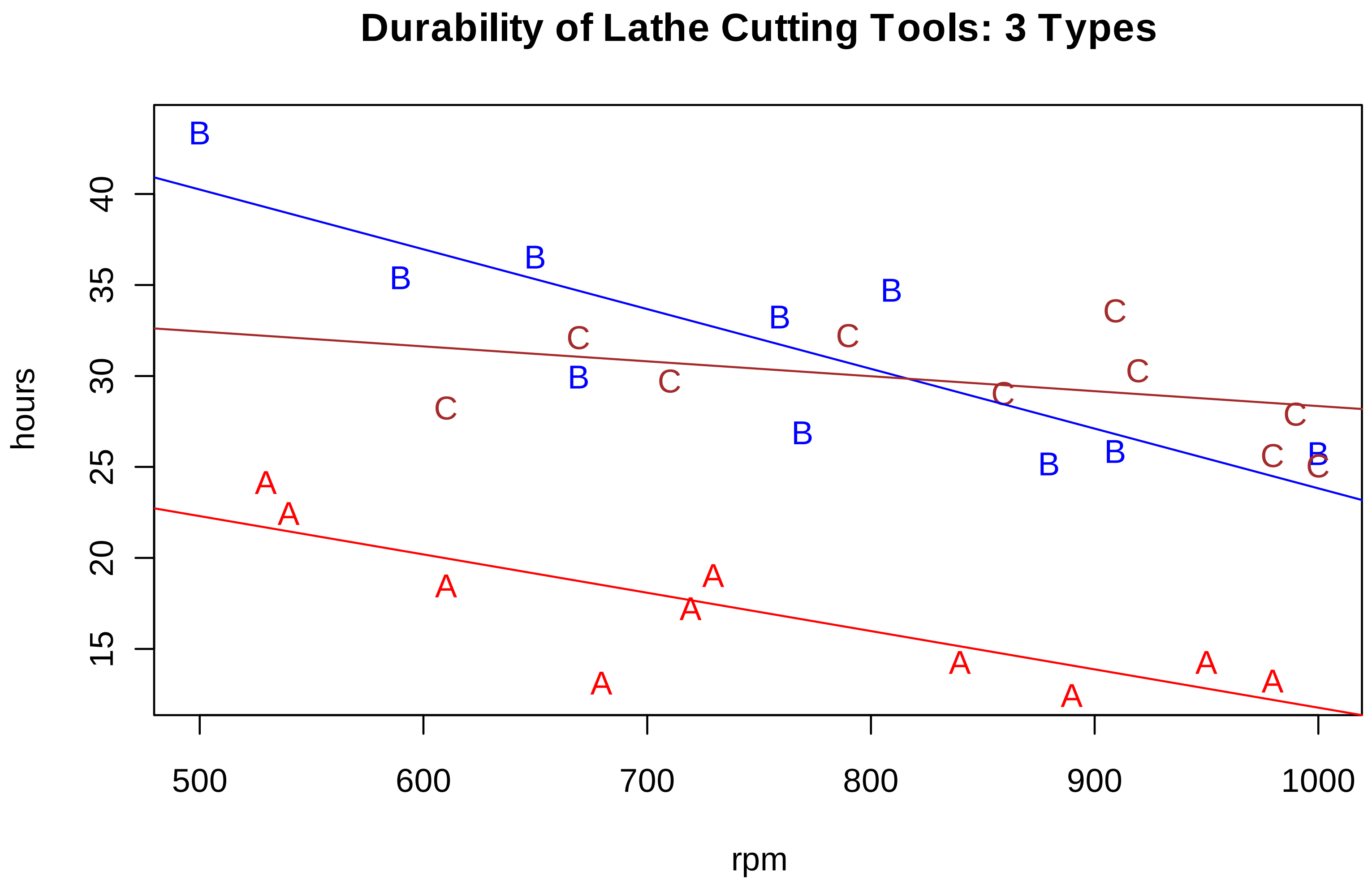
There are now 3 tool types A, B, C:

x_2	x_3	
0	0	*for observations of type A*
1	0	*for observations of type B*
0	1	*for observations of type C*

Main effect model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + E$

With interactions: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + E$

Three Types of Cutting Tools



Summary Output

```
> summary(lm(hours ~ rpm * tool, data = abc.lathe)

Coefficients:Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.774760   4.496024   7.290 1.57e-07 ***
rpm         -0.020970   0.005894  -3.558  0.00160 **
toolB       23.970593   6.568177   3.650  0.00127 **
toolC        3.803941   7.334477   0.519  0.60876
rpm:toolB   -0.011944   0.008579  -1.392  0.17664
rpm:toolC    0.012751   0.008984   1.419  0.16869
---
Residual standard error: 2.88 on 24 degrees of freedom
Multiple R-squared: 0.8906,    Adjusted R-squared: 0.8678
F-statistic: 39.08 on 5 and 24 DF,  p-value: 9.064e-11
```

This summary is of limited use for deciding about model complexity. We require hierarchical model comparisons!

Applied Statistical Regression
AS 2012 – Week 06

Inference with Categorical Predictors

Do not perform individual hypothesis tests on factors that have more than 2 levels, they are meaningless!

Question 1: do we have different slopes?

$$H_0 : \beta_4 = 0 \;\; and \;\; \beta_5 = 0 \text{ against } H_A : \beta_4 \neq 0 \;\; and/or \;\; \beta_5 \neq 0$$

Question 2: is there any difference altogether?

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \text{ against } H_A : any \; of \;\; \beta_2, \beta_3, \beta_4, \beta_5 \neq 0$$

→ Again, R provides convenient functionality: `anova()`

Applied Statistical Regression
AS 2012 – Week 06

Anova Output

```
> anova(fit.abc)

Analysis of Variance Table
          Df  Sum Sq Mean Sq F value    Pr(>F)    
rpm        1  139.08  139.08 16.7641  0.000415 ***
tool       2 1422.47  711.23 85.7321 1.174e-11 ***
rpm:tool   2   59.69   29.84  3.5974  0.043009 *  
Residuals 24  199.10    8.30
```

→ The interaction term is weakly significant. Thus, there is some weak evidence for the necessity of different slopes.

→ The p-value for the tool variable includes omitting interaction and main effect. Being strongly significant, we have strong evidence that tool type distinction is needed.

Applied Statistical Regression
AS 2012 – Week 06

Fazit über Vielfalt

Modellbeispiel zeigen

Y = x + x2 + log(x) + x1*x2 etc... (siehe vorne)

Wie entscheiden:

- Trsf. First Aid oder Modelldiagnostik

$$x' = \log(x) \qquad y' = \log(y)$$

- Interaktionen: Testen/Variablenselektion oder Modelldiagnostik

$$x' = \sqrt{x} \qquad y' = \sqrt{y}$$

Das Erkennen von Modelldefiziten und Verbesserungsmöglichkeiten unterscheidet den Profi vom Anfänger. Viele Tools stehen zur Verfügung. Wir lernen sie kennen.

$$x' = \sin^{-1}\left(\sqrt{x}\right)$$